

LARGE SCALE DISTRIBUTED DISTANCE METRIC LEARNING

Pengtao Xie & Eric Xing

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15213
{pengtaox, epxing}@cs.cmu.edu

ABSTRACT

In large scale machine learning and data mining problems with high feature dimensionality, the Euclidean distance between data points can be uninformative, and Distance Metric Learning (DML) is often desired to learn a proper similarity measure (using side information such as example data pairs being similar or dissimilar). However, high dimensionality and large volume of pairwise constraints in modern big data can lead to prohibitive computational cost for both the original DML formulation in Xing et al. (2002) and later extensions. In this paper, we present a distributed algorithm for DML, and a large-scale implementation on a parameter server architecture. Our approach builds on a parallelizable reformulation of Xing et al. (2002), and an asynchronous stochastic gradient descent optimization procedure. To our knowledge, this is the first distributed solution to DML, and we show that, on a system with 256 CPU cores, our program is able to complete a DML task on a dataset with 1 million data points, 22-thousand features, and 200 million labeled data pairs, in 15 hours; and the learned metric shows great effectiveness in properly measuring distances.

1 INTRODUCTION

Learning a proper distance metric Xing et al. (2002); Bilenko et al. (2004); Bar-Hillel et al. (2005); Globerson & Roweis (2005); Weinberger et al. (2005); Davis et al. (2007); Guillaumin et al. (2009); Kostinger et al. (2012); Mensink et al. (2012) is essential for many distance based data mining and machine learning algorithms, such as retrieval Hoi et al. (2008), k-means clustering Xing et al. (2002) and k nearest neighbor (kNN) classification Weinberger et al. (2005). A commonality of these algorithms is that their accuracy depends critically on a good distance metric M between points, especially when the dataset is high-dimensional.

As first formulated in Xing et al. (2002), mathematically, a Distance Metric Learning (DML) problem can be formulated as semi-supervised learning problem where, given side information in the form of data pairs that are determined to be similar or dissimilar, we learn a metric M , which places similar data pairs close to each other, and dissimilar data pairs as far apart as possible. This leads to a quadratic program whose size grows super-linearly with the size of the data and of the side information. Specifically, let M defines a Mahalanobis distance $(x - y)^T M (x - y)$, where x, y are d -dimensional feature vectors and $M \in \mathbb{R}^{d \times d}$ is a positive semidefinite matrix (to be learned). Because M is a d -by- d matrix, when the feature dimension d is huge — such as in web-scale problems where web pages are represented with bag-of-words (BOW) vectors that are millions of words long Ahmed et al. (2012), or in computer vision where hundreds of thousands of features are routinely extracted from images Lin et al. (2011) — the size of M quickly becomes intractable for a single machine; e.g., if d contains 1 million features, then M contains 1 trillion parameters. Storing this M requires ~ 4 terabytes of memory, to say nothing of the massive computational cost of learning so many parameters. Even worse, the number of labeled data pairs can easily exceed billions or even trillions, particularly in web data. For instance, in Flickr, photos are organized into many interests groups (e.g., tiger, sushi, car) by the users — if we simply regard photos in the same group as similar, and those in different groups as dissimilar, we can rapidly generate a huge number of similar/dissimilar pairs, as Flickr contains billions of photos and over ten million groups.

For problems involving big data volume, one would naturally speculate that modern distributed system should come for the rescue, and platforms such as Hadoop Foundation (2009) or Spark Zaharia et al. (2012) may straightforwardly offer to scale up the originally sequential algorithms intended for a single machine. But what makes DML on large-scale data nontrivial even with popular distributed platforms is that its basic formulation requires both substantial redesign of the original algorithm, and a more optimization-friendly parallel communication strategy not supported under the currently popular bulk synchronization parallelism (BSP) adopted by Hadoop and Spark. Specifically, DML is a semi-definite programming (SDP) problem Xing et al. (2002), which requires eigen-decomposition of the d -by- d Mahalanobis distance matrix M after each update. Because eigendecomposition requires $O(d^3)$ computation, it is outright infeasible when the feature dimension d is high, no matter how many machines are available. Furthermore, DML is an optimization problem with hard constraints (where each similar/dissimilar data pair corresponds to one constraint); this makes the distribution and parallel learning of parameters M very difficult, as costly, frequent inter-machine synchronization must be employed to keep the machines' local views of M consistent with each other. A BSP model would make this operation very expensive.

Motivated by these challenges, we explore and validate new approaches to performing distributed DML on large scale problems. On the algorithm side, inspired by ideas from Weinberger et al. (2005), we reformulate DML to make it tractable on high-dimensional data and amenable for distributed optimization — specifically, we re-represent M with an alternate decomposition $L^T L$, and perform optimization directly over L . This not only preserves the semi-definite property of M , but also avoids the costly $O(d^3)$ eigen-decomposition. Moreover, to solve the inter-machine parameter synchronization challenge caused by the hard DML constraints, we use slack variables to relax those constraints, and then transform the relaxed constraints into a hinge loss function. This makes the distributed optimization much easier, yet does not hurt the effectiveness of the learned distance metric, as our validation will show.

To solve this reformulated DML problem, we build a distributed system that uses asynchronous stochastic gradient descent to do parameter learning of M . Similar and dissimilar data pairs are partitioned onto different worker machines and each worker stores a local view of the parameter. At each iteration, each worker randomly picks up a mini-batch of data pairs, computes a stochastic gradient and uses the gradient to update the local parameter copy. Because the gradients computed at each worker need to be seen and utilized by other workers, we use a centralized parameter server to synchronize the parameters among workers, which has been proven to be effective both empirically and theoretically Ahmed et al. (2012); Dean et al. (2012); Ho et al. (2013); Li et al. (2014).

The major contributions of this work are summarized as follows:

- We design and implement a distributed framework to support large scale distance metric learning. To our best knowledge, this is the first work targeting the parallel and distributed metric learning on large datasets.
- We use asynchronous stochastic gradient descent to optimize DML in distributed setting and implement an efficient centralized parameter server to synchronize the parameters across machines.
- We provide an efficient and easy-to-use implementation of the framework, that we plan to open-source.
- We conducted distributed distance metric learning on large datasets and demonstrate the efficiency and effectiveness of our system.

In Section 2, we discuss related work, before introducing our reformulation of distance metric learning in Section 3. Section 4 presents our distributed framework for parallel distance metric learning, with experimental results in Section 5.

2 RELATED WORKS

Distance metric learning Xing et al. (2002); Bilenko et al. (2004); Bar-Hillel et al. (2005); Globerson & Roweis (2005); Weinberger et al. (2005); Davis et al. (2007); Guillaumin et al. (2009); Kostinger et al. (2012); Mensink et al. (2012) has been widely studied in many works. Given some data pairs labeled as similar or dissimilar, these algorithms try to learn distance metrics to make similar pairs

close to each other and separate dissimilar pairs apart. The most widely used distance metric is the Mahalanobis distance. Xing *et al* Xing et al. (2002) used semidefinite programming to learn a Mahalanobis distance metric for clustering under similarity and dissimilarity constraints. They aim to minimize the distance of similar pairs while separating dissimilar pairs with a certain margin. Weinberger *et al* Weinberger et al. (2005) and Mensink *et al* Mensink et al. (2012) employed a similar semidefinite formulation for k-nearest neighbor classification. The metric is trained with the goal that the k-nearest neighbors always belong to the same class while examples from different classes are separated by a large margin. Globerson and Roweis Globerson & Roweis (2005) proposed a formulation aiming to collapse all examples in the same class to a single point and push examples in other classes infinitely far away. Davis *et al* Davis et al. (2007) proposed information theoretic metric learning which minimizes the differential relative entropy between two multivariate Gaussians under the constraints on the distance function.

All the above-mentioned works are designed and implemented on single machine and cannot handle large scale data efficiently. Kostinger *et al* Kostinger et al. (2012) proposed a fast DML method based on likelihood-ratio test, which does not require iterative optimization procedures and hence is very efficient and scalable. However, distance metrics learned by this method yield poor performance empirically and this method is less flexible (can only deal with equivalence constraints). Our method learns effective distance metrics on multiple datasets and can flexibly deal with not only pair-wise constraints but also triple-wise constraints. In terms of the size of the problem, Mensink *et al* Mensink et al. (2012) achieved similar scale on ImageNet Deng et al. (2009) dataset as we do. However, in their work, DML was performed on a single machine, which takes days to finish. In our work, with the distributed framework, we can finish learning within 15 hours.

Designing and implementing parameter servers (PS) for distributed machine learning have been explored in several works Ahmed et al. (2012); Dean et al. (2012); Ho et al. (2013); Li et al. (2014). Most of these designs contain a centralized server and a collection of workers. The central server maintains the global parameter and each worker has a local copy of the parameter. Workers compute local updates of the parameter and push the updates to the central server. The central server aggregates updates from workers, apply the update to the global parameter and push the fresh global parameter back to workers. The existing parameter servers differ in consistency control. In Bulk Synchronous Parallel (BSP) systems like Hadoop Dean & Ghemawat (2008), Spark Zaharia et al. (2012), workers must wait for each other at the end of every iteration. In asynchronous parameter server Ahmed et al. (2012); Dean & Ghemawat (2008), all workers work on their own pace and never waits. Staleness Synchronous Parallel Ho et al. (2013) seeks a balance between BSP and ASP. Workers are allowed to see different versions of the parameter, but the difference is bounded. None of the above works has studied distance metric learning (DML) before and it is unclear whether the parameter server based framework can be effective for distributed DML.

3 DISTANCE METRIC LEARNING

In the original formulations Xing et al. (2002); Globerson & Roweis (2005); Weinberger et al. (2005) of distance metric learning, a semi-definite programming problem (SDP) needs to be solved. SDP requires eigen-decomposition of the Mahalanobis distance matrix, which is very hard to do on high dimensional data, if not impossible. To make DML scalable on large scale problems, similar to Weinberger et al. (2005), we do the reformulations in two ways. First, we factorize the Mahalanobis matrix M into $M = L^T L$ and do the optimization over L instead of M . The factorization ensures M still to be positive semi-definite, but avoids the eigen-decomposition of M . Second, we use slack variable to relax the constraints over dissimilar pairs and use hinge loss to replace the constraints. This can turn the original constrained problem into an unconstrained problem which makes its distributed optimization much easier without sacrificing the quality of learned distance metric.

3.1 REFORMULATION OF DISTANCE METRIC LEARNING

Distance metric learning is a family of algorithms and has various formulations regarding the metric to learn, the form of distance supervision and how the objective function is defined. Among them, the most popular setting is: 1, metric: Mahalanobis distance $(x - y)^T M (x - y)$, where x and y are data samples and M is a symmetric and positive semidefinite matrix to be learned; 2, the form of distance supervision: pairs of data samples either labeled as similar or dissimilar; 3, learning

objective: learn a distance metric to place similar points as close as possible and dissimilar points as far as possible. Given a set of pairs labeled as similar $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{S}|}$ and a set of pairs labeled dissimilar $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}|}$, DML learns the Mahalanobis distance by optimizing the following problem

$$\begin{aligned} \min_M \quad & \sum_{(x,y) \in \mathcal{S}} (x-y)^\top M (x-y) \\ \text{s.t.} \quad & (x-y)^\top M (x-y) \geq 1, \forall (x,y) \in \mathcal{D} \\ & M \succeq 0 \end{aligned} \quad (1)$$

where $M \succeq 0$ denotes that M is required to be positive semidefinite. This optimization problem tries to minimize the Mahalanobis distances between all pairs labeled as similar while separating dissimilar pairs with a margin 1. M is required to be positive semidefinite to ensure that Mahalanobis distance is a valid metric.

While this problem can be solved using projected gradient descent (PGD), it turns out to be infeasible for high dimensional problems. In each iteration of PGD, one first computes the gradient of M w.r.t the objective function $\sum_{(x,y) \in \mathcal{S}} (x-y)^\top M (x-y)$, then updates M using gradient descent and finally project M onto the convex set specified by the two constraints. Doing projection requires the eigen-decomposition of M , whose complexity is $O(d^3)$, where d is the feature dimension of the data points. For high dimension problems where d can be hundreds of thousands or even millions, eigen-decomposition is extremely hard and inefficient, if not impossible. Note that a symmetric and positive semidefinite matrix M can always be factorized into $M = L^\top L$ Weinberger et al. (2005), where L is a matrix of size $k \times d$ and $k \leq d$. Replacing M with $L^\top L$, the problem defined in Eq.(1) can be written as

$$\begin{aligned} \min_L \quad & \sum_{(x,y) \in \mathcal{S}} \|L(x-y)\|^2 \\ \text{s.t.} \quad & \|L(x-y)\|^2 \geq 1, \forall (x,y) \in \mathcal{D} \end{aligned} \quad (2)$$

The constraint requires the distance between dissimilar pairs to be greater than a margin of 1, which is hard to enforce in the distributed setting where communication between machines is limited. We adopt a strategy similar to Weinberger et al. (2005), which introduces slack variables ξ to relax the hard constraint in Eq.(2) and get

$$\begin{aligned} \min_L \quad & \sum_{(x,y) \in \mathcal{S}} \|L(x-y)\|^2 + \lambda \sum_{(x,y) \in \mathcal{D}} \xi_{x,y} \\ \text{s.t.} \quad & \|L(x-y)\|^2 \geq 1 - \xi_{x,y}, \xi_{x,y} \geq 0, \forall (x,y) \in \mathcal{D} \end{aligned} \quad (3)$$

Using hinge loss, the constraint in Eq.(3) can be further eliminated

$$\min_L \sum_{(x,y) \in \mathcal{S}} \|L(x-y)\|^2 + \lambda \sum_{(x,y) \in \mathcal{D}} \max(0, 1 - \|L(x-y)\|^2) \quad (4)$$

To this end, we turn the original constrained problem into an unconstrained one and the optimization becomes much easier. We use stochastic gradient method to do optimization.

4 DISTRIBUTED PARALLEL DISTANCE METRIC LEARNING

In large scale distance metric learning problems, the number of similar and dissimilar pairs can be tens of millions. The dimensionality of features can be hundreds of thousands. Learning distance metrics on such large datasets on a single machine is infeasible. To solve this problem, we design a distributed framework to do parallel DML. We use asynchronous stochastic gradient descent to do the optimization. We partition the similar pairs and dissimilar pairs onto different machines. Each machine uses the data pairs it holds to update the model parameters in an asynchronous manner. Parameters on different machines are synchronized through a centralized parameter server. In terms of the form of distance supervision, we focus on pair-wise constraints (e.g., i is similar to j ; j is dissimilar to k) in this work. Our framework can be easily extended to support triple-wise constraints Weinberger et al. (2005) (e.g., i is more similar to j than to k).

4.1 PARAMETER SYNCHRONIZATION

To learn L in a distributed environment with P worker machines, we partition the similarity pair \mathcal{S} and dissimilar pair \mathcal{D} into P pieces $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_P$ and $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_P$ and each machine holds one

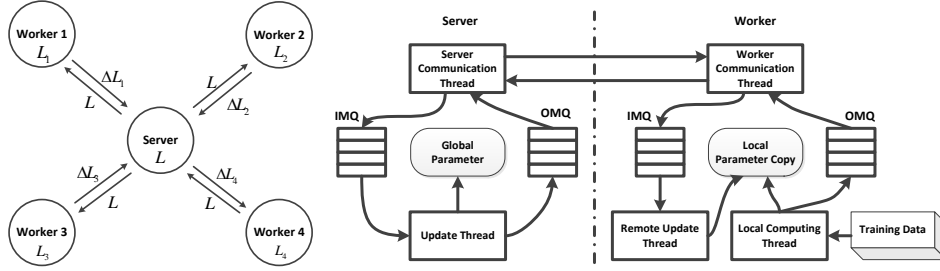


Figure 1: Logical View of Parameter Server. A centralized server maintains the global parameter L . Each work p has a local copy L_p of the global parameter. In each iteration, each work takes a data pair and computes a gradient update ΔL_p and sends ΔL_p to the central server. The server aggregates the gradients received from workers and uses them to update the global parameter L . L is then pushed back to each worker and workers replace their local copy with L .

piece. Each machine p has a local copy L_p of the global parameter L . And different parameter copies are synchronized across machines to ensure they are as much the same as possible. Similar to Ahmed et al. (2012); Dean & Ghemawat (2008); Ho et al. (2013); Lee et al. (2014), we use a centralized parameter servers to synchronize these parameter copies. The central server holds the global parameter L . Each worker communicates with the central server and workers do not communicate with each other. At each iteration of the stochastic gradient descent algorithm, each worker p randomly samples a minibatch of data pairs from both the similar pair set \mathcal{S}_p and the dissimilar pair set \mathcal{D}_p it holds and computes a gradient update ΔL_p using the local parameter copy L_p and the minibatch. Worker p sends ΔL to the central server. The central server aggregates updates received from workers and use them to update the central parameter L . The central server sends the updated L to each worker and the worker replaces its local parameter copy L_p with L .

4.2 IMPLEMENTATION DETAILS

The centralized parameter server contains one central server and P workers. Workers communicate with the central server and they do not communicate with each other. The server maintains the global parameter and each worker has a local copy of the parameter. On the server side, there are two threads: 1) update thread; 2) communication thread. The server maintains two message queues to exchange messages with workers: 1) inbound message queue; 2) outbound message queue. The communication thread receives gradient updates from workers and puts them into the inbound message queue. The update thread takes a batch of gradient updates from the inbound message queue and uses them to update the central parameter. Then the update thread puts the updated central parameter to the outbound message queue. The communication thread takes updated parameter out of the outbound message queue and sends it to all workers.

On the worker side, there are three threads: 1) local computing thread; 2) remote update thread; 3) communication thread. The worker also maintains two message queues: 1) inbound message queue; 2) outbound message queue. At each iteration, the local computing thread takes a minibatch of data pairs, computes the gradient, uses the gradient to update the local parameter copy and puts the gradient into the outbound message queue. The communication thread sends the gradients in the outbound message queue to the server. On the other hand, it receives fresh parameters from the server and puts them into the inbound message queue. The remote update thread takes parameters out of the inbound message queue and uses them to replace the local parameter copy on this worker.

The server threads and worker threads execute in a best-effort manner, which means whenever some work is available to do, the threads do it without caring about other threads. For example, whenever there is a message in the outbound message queue, the communication thread sends them out. Whenever there is a message in the inbound message queue, the update thread uses it to update the parameter. Each thread behaves as if no other threads exist. The threads are coordinated indirectly by the message queues.

Table 1: Statistics of Datasets

Dataset	feat. dim	k	# parameters	#samples	#similar pairs	#dissimilar pairs
MNIST	780	600	0.47M	60K	100K	100K
ImNet-60K	21504	10000	220M	63K	100K	100K
ImNet-1M	21504	1000	21.5M	1M	100M	100M

5 EXPERIMENTS

5.1 DATASETS

We use three datasets in our experiments. Table 1 summarizes the statistics of these datasets. The first one is the MNIST handwritten digits LeCun et al. (1998). It has 60K training images and 10K testing images. Each image is of size 32×32 and has a digit label. Images are represented with raw pixels, with a dimensionality of 780. We randomly sample 100K “similar” pairs and 100K “dissimilar” pairs from the training images. If two images are from the same digit, we label them as “similar”. If two images are from different digits, we label them as “dissimilar”.

The second dataset we use is ImageNet-63K, which has 63K training images randomly selected from the ImageNet Deng et al. (2009) dataset. Each image is associated with a class label and the total number of classes is 1000. Images are represented with Locality-constrained Linear Coding (LLC) Wang et al. (2010), with a dimensionality of 21504. We randomly sample 100K “similar” pairs and 100K “dissimilar” pairs from the training images. If two images are from the same class, we label them as “similar”. If two images are from different classes, we label them as “dissimilar”.

The third dataset is ImageNet-1M. It has 1 million training images and 63K testing images randomly selected from ImageNet Deng et al. (2009). The total number of classes is 1000. Images are represented with LLC features. We randomly sample 100M “similar” pairs and 100M “dissimilar” pairs to learn the distance metric.

5.2 EXPERIMENTAL SETUP

In all the experiments, the tradeoff parameter λ is set to 1 and the threshold c is set to 1. For MNIST, we set k (the number of rows of L) to 600. At each iteration, we use a minibatch of 1000 pairs (500 similar pairs and 500 dissimilar pairs) to do the update. For ImageNet-63K, we set k to 10000. The number of model parameters is about 220 million. At each iteration, we use a mini-batch of 100 pairs (50 similar pairs and 50 dissimilar pairs) to do the update. For ImageNet-1M, we set k to 1000 to make the computation tractable. The mini-batch size is 1000 (500 similar pairs and 500 dissimilar pairs). Experiments on MNIST are done on machines each of which has 16 CPUs and 64G main memory. Experiments on ImageNet-63K and ImageNet-1M are performed on machines each of which has 64 CPUs and 128G main memory.

5.3 SCALABILITY OF OUR FRAMEWORK

We evaluated the scalability of our framework w.r.t the number of CPU cores on three datasets. Figure 2(a) shows the convergence curves on MNIST dataset under different number of CPU cores. The horizontal axis corresponds to running time measured in minutes. The vertical axis corresponds to objective values. Figure 2(b) and 2(c) show the convergence curves on ImageNet-63K and ImageNet-1M respectively. As can be seen from the three figures, increasing the number of machines consistently increases the convergence speed.

To measure how our framework can speedup the training as we increase the number of machines (cores), for each machine setting we record the running time that the objective value is decreased to p , where p is the objective value achieved by one single machine at the end of training. The speedup factor of n machines is calculated as t_n/t_1 , where t_n is the running time of n machines and t_1 is the runtime of 1 machine. Figure 3(a) shows the speedup factors on MNIST dataset. The horizontal axis corresponds to the number of cores. The vertical axis shows speed up factors. The blue curve is linear speedup (optimal speedup). The red curve shows the speed up factors achieved by our framework. As can be seen from this figure, our framework achieves near optimal speedup

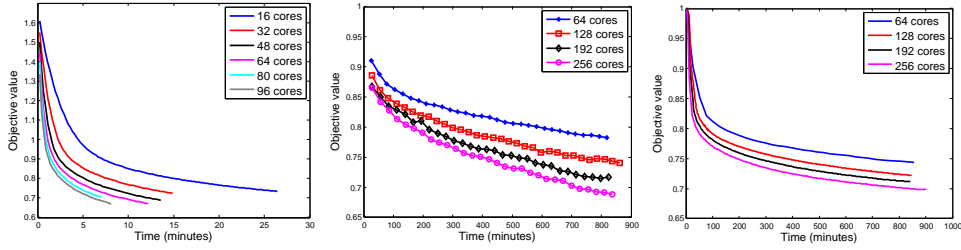


Figure 2: (a) Convergence curves on MNIST dataset. (b) Convergence curves on ImageNet-63K dataset. (c) Convergence curves on ImageNet-1M dataset.

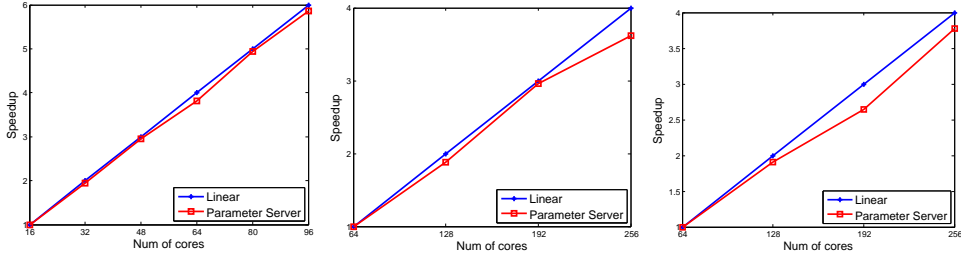


Figure 3: (a) Speedup on MNIST dataset. (b) Speedup on ImageNet-63K dataset. (c) Speedup on ImageNet-1M dataset.

under all machine settings with different CPU cores. Figure 3(b) and 3(c) shows the speedup on ImageNet-63K and ImageNet-1M dataset respectively. Our framework achieves speedups which are very close to linear under different number of cores. With 4 machines (256 cores in total), our framework achieves 3.6 times speedup on ImageNet-63K dataset and achieves 3.8 times speedup on ImageNet-1M dataset compared with those on 1 machine (16 cores in total). This demonstrates that our framework scales very well with the number of CPU cores (machines).

The scalability of our framework attributes to the sufficient use of CPU cores to do computation and efficient parameter synchronization among workers. We distribute the computation over all available machines and allocate a copy of the model parameters to each machine. Computing threads on each machine always have access to the parameters through the parameter copy, thereby, computation can be ceaselessly performed without blocking caused by the coordination and communication with other machines. Each machine runs as if no other machines exist. This ensures that all CPU cores can be sufficiently used. On the other hand, we synchronize the parameter copies of different workers using a centralized server, to ensure that each work’s update can be timely contributed to the global parameters and the computing threads on each worker can use the most-up-to-date parameters to do computation. The synchronization happens in background and does not require blocking computation.

5.4 QUALITY OF THE LEARNED DISTANCE METRIC

We also measure the quality of the learned distance metric. We compare our reformulation in Eq.(4) with the original formulation Xing et al. (2002) in Eq.(1) (denoted by Xing2012) and with Information Theoretical Metric Learning (ITML) Davis et al. (2007) and the likelihood test based method (KISS) Kostinger et al. (2012). All methods (including our method) are implemented with MATLAB and runs on a single thread. For each method, we learn a distance metric on the MNIST training set. In ITML, we set the tradeoff parameter γ is set to 0.001. In KISS, the feature dimension is reduced to 600 using Principal Component Analysis (PCA) to ensure the covariance matrices are invertible.

To evaluate the effectiveness of the learned metric, we randomly sample 10K similar pairs and 10K dissimilar pairs from 100K held-out testing images and use the metric to judge whether these pairs are similar or dissimilar. If the distance is great than some threshold t , the pair is regarded as similar. Otherwise, the pair is regarded as dissimilar. We use two evaluation metrics: average precision and precision-recall curves.

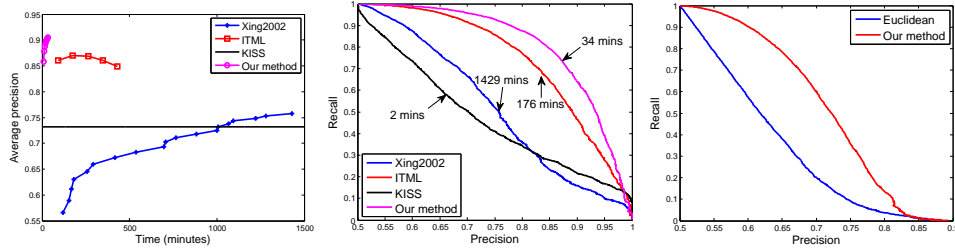


Figure 4: (a) Average precision versus running time on MNIST dataset. (b) Precision-recall curves on MNIST dataset. (c) Precision-recall curves on ImageNet-1M dataset.

Figure 4(a) shows the average precision versus running time. Figure 4(b) shows the best precision-recall curve achieved by each method. As can be seen from the two figures, our method is not only much faster, but also achieves better performance. Optimizing the original formulation in Eq.(1) (Xing2002) requires solving a linear constrained least square (LCLS) problem at each iteration, which is very costly. On the MNIST dataset, the number of variables to be optimized in LCLS is 680K and the number of constraints is 100K. Solving such a LCLS problem in each iteration is highly demanding. Besides, the original formulation requires eigen-decomposition of the Mahalanobis matrix M , which is very costly when the feature dimension is high. In ITML, the computational complexity on each data pairs is $O(d^2)$, where d is the feature dimension. In our method, the complexity is $O(dk)$, where k is usually smaller than d . From Figure 4(a), we observe that in ITML, the precision is not consistently increasing as running time increases. This is because in each iteration of ITML, a single data pair is used to update the parameter, which may incur high variance. It is unclear how to extent ITML to mini-batch update to make the algorithm stable. KISS computes a distance metric in one shot and does not require iterative optimization, thereby, it is very fast. It learned a metric in 2 minutes. However, the metric yields very poor performance. The average precision is only 0.73 (Figure 4(a)) and the precision-recall curve (Figure 4(b)) is much worse than other methods. Our method is very efficient. The training is finished in about half a hour on a single thread while Xing2002 takes about 24 hours and ITML takes about 3 hours. With our distributed framework, the speed can be further improved significantly as shown in Figure 3(a). In addition, our method is very effective. It achieves an average precision of 0.90, which cannot be achieved by the other three methods.

We also evaluate the effectiveness of the distance metric learned on ImageNet-1M. From 63K held-out testing images, we randomly sample 100K similar pairs and 100K dissimilar pairs and compute precision-recall curves, which are shown in Figure 4(c). The blue curve is obtained from Euclidean distance computed on original feature vectors. The red curve corresponds to Mahalanobis distance computed under the learned distance metric. As can be seen from the figure, with distance metric learning, the performance is greatly improved.

6 CONCLUSIONS

In this paper, we reformulate the original DML problem into an unconstrained optimization problem that is amenable for parallelization, and use it as the basis for a distributed framework to support large scale distance metric learning. Our framework makes use of asynchronous stochastic gradient descent for distributed optimization, where the computation is distributed across worker machines, which use a centralized parameter server to synchronize parameters between them. Experiments on three datasets demonstrate the efficiency and effectiveness of our framework, at previously-unseen problem scales with better accuracy than previous methods. We believe our work is the first to address distance metric learning at such large scales, in an intelligent distributed fashion.

REFERENCES

- Ahmed, Amr, Aly, Moahmed, Gonzalez, Joseph, Narayanamurthy, Shravan, and Smola, Alexander J. Scalable inference in latent variable models. In *WSDM*, pp. 123–132. ACM, 2012.
- Bar-Hillel, Aharon, Hertz, Tomer, Shental, Noam, and Weinshall, Daphna. Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, 6(6):937–965, 2005.

- Bilenko, Mikhail, Basu, Sugato, and Mooney, Raymond J. Integrating constraints and metric learning in semi-supervised clustering. In *ICML*, pp. 11. ACM, 2004.
- Davis, Jason V, Kulis, Brian, Jain, Prateek, Sra, Suvrit, and Dhillon, Inderjit S. Information-theoretic metric learning. In *ICML*, pp. 209–216. ACM, 2007.
- Dean, Jeffrey and Ghemawat, Sanjay. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.
- Dean, Jeffrey, Corrado, Greg, Monga, Rajat, Chen, Kai, Devin, Matthieu, Mao, Mark, Senior, Andrew, Tucker, Paul, Yang, Ke, Le, Quoc V, et al. Large scale distributed deep networks. In *NIPS*, pp. 1223–1231, 2012.
- Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, and Fei-Fei, Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. IEEE, 2009.
- Foundation, The Apache Software. Apache hadoop. In <http://hadoop.apache.org/core/>, 2009.
- Globerson, Amir and Roweis, Sam T. Metric learning by collapsing classes. In *NIPS*, pp. 451–458, 2005.
- Guillaumin, Matthieu, Verbeek, Jakob, and Schmid, Cordelia. Is that you? metric learning approaches for face identification. In *ICCV*, pp. 498–505. IEEE, 2009.
- Ho, Qirong, Cipar, James, Cui, Henggang, Lee, Seunghak, Kim, Jin Kyu, Gibbons, Phillip B, Gibson, Garth A, Ganger, Greg, and Xing, Eric. More effective distributed ml via a stale synchronous parallel parameter server. In *NIPS*, pp. 1223–1231, 2013.
- Hoi, Steven CH, Liu, Wei, and Chang, Shih-Fu. Semi-supervised distance metric learning for collaborative image retrieval. In *CVPR*, pp. 1–7. IEEE, 2008.
- Kostinger, M, Hirzer, Martin, Wohlhart, Paul, Roth, Peter M, and Bischof, Horst. Large scale metric learning from equivalence constraints. In *CVPR*, pp. 2288–2295. IEEE, 2012.
- LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- Lee, Seunghak, Kim, Jin Kyu, Zheng, Xun, Ho, Qirong, Gibson, Garth A, and Xing, Eric P. Primitives for dynamic big model parallelism. *arXiv preprint arXiv:1406.4580*, 2014.
- Li, Mu, Andersen, David G, Park, Jun Woo, Smola, Alexander J, Ahmed, Amr, Josifovski, Vanja, Long, James, Shekita, Eugene J, and Su, Bor-Yiing. Scaling distributed machine learning with the parameter server. In *Proc. OSDI*, pp. 583–598, 2014.
- Lin, Yuanqing, Lv, Fengjun, Zhu, Shenghuo, Yang, Ming, Cour, Timothee, Yu, Kai, Cao, Lian-gliang, and Huang, Thomas. Large-scale image classification: fast feature extraction and svm training. In *CVPR*, pp. 1689–1696. IEEE, 2011.
- Mensink, Thomas, Verbeek, Jakob, Perronnin, Florent, and Csurka, Gabriela. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *ECCV*, pp. 488–501. Springer, 2012.
- Wang, Jinjun, Yang, Jianchao, Yu, Kai, Lv, Fengjun, Huang, Thomas, and Gong, Yihong. Locality-constrained linear coding for image classification. In *CVPR*, pp. 3360–3367. IEEE, 2010.
- Weinberger, Kilian Q, Blitzer, John, and Saul, Lawrence K. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, pp. 1473–1480, 2005.
- Xing, Eric P, Jordan, Michael I, Russell, Stuart, and Ng, Andrew Y. Distance metric learning with application to clustering with side-information. In *NIPS*, pp. 505–512, 2002.
- Zaharia, Matei, Chowdhury, Mosharaf, Das, Tathagata, Dave, Ankur, Ma, Justin, McCauley, Murphy, Franklin, Michael J, Shenker, Scott, and Stoica, Ion. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pp. 2–2. USENIX Association, 2012.